DOCUMENT RESUME

ED 232 556 HE 016 477

AUTHOR Follman, John
TITLE Student Ratings of Faculty Teaching Effectiveness: Revisited. ASHE 1983 Annual Meeting Paper.
PUB DATE Mar 83
NOTE 17p.; Paper presented at the Annual Meeting of the Association for the Study of Higher Education (Washington, DC, March 25-26, 1983).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; *Faculty Evaluation; Higher Education; *Rating Scales; Reliability; Stereotypes; *Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Trend Analysis; Validity
IDENTIFIERS *ASHE Annual Meeting

ABSTRACT

The literature from 1973-1983 on student ratings of faculty teaching effectiveness is reviewed. Among the consistent findings over the decade are the following: high student rating scale reliability, the limited influence of instructor and student demographic variables on student ratings of teacher effectiveness, and faculty perceptions that evaluation of their performance is important. Myths that continue to be identified in the literature include the following: the belief that student raters do not discriminate among items and that they are insufficiently knowledgeable and/or sophisticated to properly evaluate their instructors, and the view that students are seduced by instructors' expressive personalities even if the instructors present little knowledge. An interesting finding is that when students are asked to list by name the best and worst instructors they have had, 15-20 percent of the instructors are on both lists. Issues remaining to be resolved are the correlation of .43 between student ratings and student achievement, the purpose of student ratings, and identifying the factors represented within rating scales. Information is included on new terms, new rating scale acronyms, new methodologies and statistics and/or applications. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED232556

STUDENT RATINGS OF FACULTY TEACHING EFFECTIVENESS: REVISITED

John Follman
University of South Florida

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John Follman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

NE 016 477

ASHE Association for the Study of Higher Education
The George Washington University/One Dupont Circle, Suite 630/Washington, D.C. 20036
(202) 296-2597

This paper was presented at the Annual Meeting of the Association for the Study of Higher Education held at the Washington Hilton in Washington, D.C. March 25-26, 1983. This paper was reviewed by ASHE and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC collection of ASHE conference papers.

Student Ratings of Faculty Teaching Effectiveness: Revisited

John Follman
University of South Florida

The author, after a 10 year hiatus, recently returned to researching effectiveness of teaching of higher education faculty especially use of student ratings. The method of return has been a comprehensive review of the relevant empirical evidence reported since 1973. During the course of this update the author perceived that he had a perspective on research progress rarely realized. This is an initial thorough review of an important literature, followed by a 10 year departure from that literature, followed by a return to that literature. Ordinarily, one's perspective tends to be limited and narrowed by his immediate interests. The author believes that the absence and subsequent return to this vast literature provides a perspective of progress not often achieved.

The initial insight of this overview is the sheer size of the literature, the empirical literature. The best estimate is about 10,000 relevant articles (for a compilation of relevant references see Follman 1982 a). If the opinion, normative, hortatory, and theoretical literature is also included, the reviewer would soon reach an asymptote of diminishing returns, especially since there is a 75 year history of empirical research, much conducted and more stimulated by the grand old man of student ratings, Professor H. H. Remmers. The author would not be surprised if the teaching effectiveness literature, itself, was as large as that associated with some of the newer "disciplines" in education such as foreign language education, mathematics education, science education, etc.

The taxonomy, developed by the author to organize his perceptions of the literature over the last 10 years, includes verities, non-verities

(myths), nonverities-changes, intrinsically interesting items, central issues, levities, new terms, new methodologies and statistics, aves, bananas, and kudos, and finally, prognoses.

Verities

Consistent findings continuing to obtain over the decade include high student rating scale reliability.

Validity of student rating scales is also a verity. In the early 70's the author reviewed the research studies of the relationship between student ratings and student achievement and found an average r of .43 (Follman 1974). Ten years later Cohen (1981) reported an average r of .43 between the same two variables.

Little research and little difference in the limited influences of instructor presage and process variables on student ratings and/or achievement was observed over the decade (Follman 1982 a).

Another little changed finding was that feedback to instructors from student ratings, especially feedback alone, has limited salutary effect on subsequent instructor pedagogical behavior (Follman 1982 a).

Another consistency is the limited influence of instructor demographic variables such as age, sex, rank, years of teaching experience, and also of student demographic variables, such as sex, grade, class level, class size, required vs elective, discipline, etc., on student ratings of faculty teaching effectiveness. On the other hand, student's expected course grade

correlates up to about .30 with student ratings and predicts as much of student rating variance as the other student demographic variables combined. Apparently students who receive lower grades than they expect rate punitively while students who receive higher grades than they expect rate benevolently (Happy Birthday Effect, Kennedy 1975). Also, students who receive high grades tend to have an internal locus of control while students who receive low grades tend to attribute their performance to their instructor (Snyder and Clair 1976).

Another verity is that faculty still perceive evaluation of their performance to be very important. Institutions contemplating heavy weighting of student ratings in faculty evaluation should employ faculty participation and proceed in all deliberation (Follman 1982 b, 1982 c).

An emerging consistency is the relative influence of course vs instructor on student ratings. The returns are not all in yet but it appears that the instructor has at least twice as much influence as the course (Follman 1982a).

Non-Verities (Myths)

One myth which has been reified is that student raters either cannot or do not discriminate among items. They do (Follman 1982 a).

A variation of this cant is that students are insufficiently knowledgeable and/or sophisticated to properly evaluate their instructors, and that ratings of students should not be collected until a year or two

later and/or they have graduated. In fact, the correlations within student ratings taken during a course and taken again a year or two later are usually in the high .70's (Follman 1982 a).

A third version of this myth is that high student ratings can be obtained by professors by pandering to the students, lightening their overload, inflating their grades, etc. There is not much evidence that much of this actually happens. The best evidence is the median r of .14 between student ratings of the teaching effectiveness of their instructors and their grades (Aleamoni 1981). There is some, other, evidence that some pandering happens (Follman 1982 c). The basic desideratum is that the r between student ratings and student achievement remains around .43, hardly overwhelming evidence that student ratings are determined by the grades students receive.

Still another version of the student rating inadequacy assertion is that students are seduced by instructors' expressive personalities even if the instructors present little or no substantive knowledge. Students do discriminate between instructors' theatrical ability and their substantive knowledge. The considerable Dr. Foxx literature, probably a cul de sac as far as the effects of instructors' personalities on achievement are concerned, demonstrates this (Abrami, Leventhal and Perry 1982).

Non-Verities (Changes)

The author detected no major issues whose underlying preponderance of evidence changed materially over the decade. It is noted though that the limited emphasis on pedagogical process in higher education has diminished.

Intrinsically Interesting Items

The most interesting finding I have observed over all the teaching effectiveness literature over the years is the three study finding that, when students are asked to list, by name, the best instructor they have had, and the worst instructor they have had, 15 to 20% of the instructors are on both lists, both lists, believe it or not (Follman 1982 a). Socrates to some students is apparently hemlock to others. If there is any one single finding that circumscribes the ceiling on the validity of the relation between student ratings and student achievement it is this.

Another interesting finding helping to account for the validity ceiling is the varied and often low correlations among the four criterion groups of teacher effectiveness, instructors themselves, peers, administrators, and students, and also with student achievement (Follman 1982 a). Inferentially, the reason for the lack of consistent, high correlations, among these groups is that each group apparently emphasizes and/or evaluates different variables.

Another interesting item is associated with students' expected course grade. Students' expected grades and their ratings of their instructors correlate between .10 and .30 generally, with higher r's for total class means than for individual students. Students rate their instructors higher the lower their grade expectancies are, and also when they receive higher grades than they expect. They rate lower when their expectancies are higher than the grades they receive. The relation between students' ratings and their expected grades is twice as high as the relation between their ratings and their actual grades (Follman 1982 a).

Central Issues

Basic issues remaining to be resolved empirically or deductively include validity, will the correlation of .43 between student ratings and student achievement increase? I do not believe that it will increase much in the forseeable future, if ever.

One approach that may strengthen this r is profile analysis, i.e., identification of portions of variance associated with subgroups within a rating group or class which probably now wash out each other's effects (Weber and Frary 1981).

Another approach is intensive research into the idiosyncrasies, anticipations, expectations, predispositions, rating tendencies, etc., of the raters which may influence some of their ratings as much as the instructor purportedly being rated (Follman 1975).

Perhaps an algebraic rationale making reliability equivalent to validity will be devised so one will not have to square (lower) the validity coefficient. Generalizability theory may be the vehicle (Aubrecht 1981, Smith 1979).

Rater training may be efficacious in reducing rater leniency and halo effect and thus increase reliability and validity. Substantial additional research is required.

A second issue remaining unresolved is the purpose of the student ratings, instructor self improvement, administrative use, or student use.

The author is completing a review of the limited literature on this issue and while it is clear that most institutions of higher learning are using student ratings, and using them administratively, their processes and weightings are nearly always implicitly, at best, articulated (Follman 1982 c).

A third unresolved issue is the number of factors represented within rating scales. A continuing problem has been to break down the well known rating scale halo effect (Follman 1982 a). The McKeachie factors of Skill, Structure, Rapport, Work Load, and Interaction (McKeachie 1979) have been replicated in many studies. These factors probably can be reproduced in most colleges where psychometric expertise obtains. Also, Marsh (Marsh, Overall and Kesler 1979) has developed a rating scale at the University of Southern California which has a number of different dimensions with factorial integrity. However the issue, analogous to the number of factors in scholastic aptitude tests, remains unresolved.

A development that would be useful in helping resolve some of these isues is a data bank of student ratings available to researchers without access to large scale student rating data.

Levities

My most amusing experience with evaluation of instruction occurred about 10 years ago when the '60's belatedly arrived on our campus. Among the baggage, besides the laid-back attitude, and drugs, was something less than scholarly fervor. On one occasion I was able to obtain a very large

class in an auditorium in which to run a rating formats variation experiment. During my orientation to the students I advised them that I was conducting an experiment on student rating scale formats in that course but that the students should rate that instructor and course just as if there were no experiment. I then distributed the rating formats, answer sheets, and electronic pencils. After the students completed and returned the forms, answer sheets, and pencils, and had departed, I policed the rows looking for any answer sheets and pencils not returned. In addition to pencils and answer sheets, I found two students sleeping (one on the floor), one lingerie item, braces, and two reefers. Concurrently, over the entire class period, a group of 8-10 students stood between the stage and the first row, stoned, talking throughout my participation and also throughout the instructor's presentation. Later, as I collated the answer sheets, I found written on one of them, in addition to the usual inanities and obscenities, the comment "I didn't know this was a course, I just stop in here to rest after selling my blood."

New Terms

Animadversion Error refers to criticism prompted by prejudice such as may occur when subordinates rate their superiors (Schuh and Crivelli 1973).

Behavioral Expectation Scales (BES) represent a type of student developed rating scale refined into critical incidents retranslated into rating dimensions (Bernardin 1977).

Cusp Catastrophe response is a mathematical model for discontinuous data applied to student ratings of faculty teaching effectiveness (Staman

1982).

Philetic teaching refers to mutual affection between instructor and students (Meredith and Ogasawara 1981).

Zetetics as a system of knowledge refers to the collecting and systematizing data on the activity of research and creative activity (Short 1973).

Student Rating Scale Acronyms

Course faculty Instrument (CFI) is a student rating scale developed by Stumpf, Freedman and Krieger (1979).

Student evaluation of teaching (SET) is the term used commonly in economics (Villard 1973).

Teacher Ratings Forms (TRF) is a scale created by Leventhal (1975).

New Methodologies and Statistics and/or Applications

Generalizability theory is the use of analysis of variance to estimate the amounts of variance associated with different factors, in this application, instructor effect, course effect and student effect and/or item effect (Aubrecht 1981, Smith 1979).

Meta-Analysis refers to the analysis of analyses, specifically integration of findings of different studies in reviews of the literature. Outcomes and features of the different studies are characterised in quantitative or quasi-quantitative terms. Finally, multivariate statistics are used to describe the outcomes and relate characteristics of the studies to the outcomes (Glass 1976).

Path Analysis is a method of arranging causal linkage or modelling using different kinds of correlations and regression analysis to demonstrate the direct and indirect effects of the independent variables Feldman (1978).

Profile Uniqueness is an approach to identify subgroups of students within a class which rate differently across items, by clustering these sub groups of raters who give the same or smilar ratings across all items of a scale. This procedue is a refinement over conventional mean statistics which mask by evening out differences (Weber and Frary 1981). This approach needs replication and refinement including a canned computer program.

Aves, Bananas, and Kudos

Aves go to the disciplines of economics and sociology which have joined old standbys psychology, educational psychology, educational research and other disciplines, in contributing substantial research to evaluation of instruction. The regression approach in many of the economists' studies has added a different dimension, i.e., monetary value of learning, ratings, publications, papers, grants, etc.

The Banana Award to Dr. Alfred E. Kahn who, some years ago chaired a committee on the Quality of Undergraduate Instruction at Cornell University (Rayder 1968).

Kudos to Ken Feldman for his classic scholarly reviews, Herbert Marsh of Southern California and Down Under for his factor analyses, and to Lawrence Aleamoni, John Centra, W. J. McKeachie, and others for their continuing contributions.

Prognoses

I do not expect the correlation between student ratings and student achievement to appreciate much in the forseeable future, if ever.

It is recommended that systematic evaluation of instructors involve teachers (selves), peers, administrators, and students, as well as student achievement. These criteria are probably all necessary because of their inconsistent and sometimes low intercorrelations. I do not expect this large scale systematic evaluation to happen but I do expect some, limited, progress in this direction.

It is also recommended that only two levels of instructor quality, superior and inferior, with the vast majority of instructors in between, be differentiated, since the differences among faculty in the middle are so small that format variations alone may be greater (Follman 1982 a).

I do expect that institutions of higher learning will become considerably more systematic about the process of collecting student ratings so that at the least, no instructor will participate in the collection of data of evaluation of himself, a clear cut conflict of interest.

BIBLIOGRAPHY

Abrami, Philip C., Leventhal, Les, and Perry, Raymond P. Educational Seduction. American Educational Research Association, New York City, March 1982.

Aleamoni, Lawrence M. Description of IRAD's charge. Note to the faculty, University of Arizona, No. 11, 1981.

Aubrecht, Judith D. Reliability, validity, and generalizability of student ratings of instruction. IDEA PAPER No. 6, Center for Faculty Evaluation and Development, Kansas State University, November, 1981.

Baker, F. B. Numerical taxonomy for educational researchers. Review of Educational Research, 1972, 42, 345-358.

Bernardin, H. John. Behavioral expectation scales versus summated scales: a fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Cohen, Peter A. Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies. Review of Educational Research, 1981, 51, 281-309.

Feldman, Kenneth A. Course characteristics and college students' ratings of their teachers: What we know and what we don't. Research in Higher Education, 1978, 9, 199-242.

Follman, John. Student ratings and student achievement. Journal Supplement Abstract Services, American Psychological Association, 1974, 4, 136.

Follman, John. Student ratings of faculty teaching effectiveness: rater or ratee characteristics. Research in Higher Education, 1975, 3, 155-167.

Follman, John. Basic questions re evaluation of college teaching. Pennsylvania-New Jersey Educational Research Associations, Philadelphia, May 17, 1982, a.

Follman, John. Evidence re evaluation of faculty for tenure, promotion, and raises. Invitational Symposium on Higher Education, Kent State University, April 23, 1982, b.

Follman, John. Recommendations re use of student ratings of teaching effectiveness. American Educational Research Association, Special Interest Group Community College Research, Madison, Wisconsin, July 16, 1982, c.

Glass, Gene V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Kennedy, W. Robert. Grades expected and grades received their relationship to students' evaluations of faculty performance. Journal of Educational Psychology, 1975, 67, 109-115.

Leventhal, Les. Teacher rating forms: critique and reformulation of previous validation designs. Canadian Psychological Review, 1975, 16, 269-276.

Marsh, Herbert W., Overall, Jesse U., and Kesler, Steven P. Class size, students' evaluations, and instructional effectiveness. American Educational Research Journal, 1979, 16, 57-69.

McKeachie, Wilbert J. Student ratings of faculty: a reprise. Academe, 1979, 65, 384-397.

Meredith, Gerald M., and Ogasawara, Todd H. Lecture size and students' ratings of instructional effectiveness. Perceptual and Motor Skills, 1981, 52, 353-353.

Peterson, Robert A., Kerin, Roger A., and Martin, Warren S. Teaching effectiveness in the decision sciences: the use of student evaluations, 1978.

Rayder, Nicholas F. College student ratings of instructors. The Journal of Experimental Education, 1968, 37, 76-81.

Schuh, Allen J., and Crivelli, Michael A. Animadversion error in student evaluations of faculty teaching effectiveness. Journal of Applied Psychology, 1973, 58, 259-260.

Short, Edmund C. Knowledge production and utilization in curriculum: a special case of the general phenomenon. Review of Educational Reserch, 1973, 43, 237-301.

Smith, Philip L. The stability of teacher performance in the same course over time. Research in Higher Education, 1979, 11, 153-165.

Staman, E. Michael. Catastrophe theory in higher education research. Research in Higher Education, 1982, 16, 41-53.

Stumpf, Stephen A., Freedman, Richard D., and Krieger, Kenneth M. Validity extension of the course-faculty instrument (CFI). Research in Higher Education, 1979, 11, 13-22.

Snyder, C. R., and Clair, Mark. Effects of expected and obtained grades on teacher evaluation and attribution of performance. Journal of Educational Psychology, 1976, 68, 75-82.

Villard, Henry H. Some reflections on student evaluation of teaching. The Journal of Economic Education, 1973, 5, 47-50.

Weber, Larry J., and Frary, Robert B. Profile uniqueness in student ratings of instruction. American Educational Research Association, Los Angeles, April 13-17, 1981.

CPT:A.F1-4